Optimizing a metabarcoding primer portfolio for species-level detection of taxa in complex mixtures of diverse fishes

Diana Baetscher[1], Nicolas Locatelli[1], Eugene Won[2], Timothy Fitzgerald[3], Peter McIntyre[1], and Nina Overgaard Therkildsen[1]

[1]Cornell University College of Agriculture and Life Sciences
[2]Cornell University
[3]Environmental Defense Fund Washington DC

December 4, 2021

Abstract

DNA metabarcoding is used to enumerate and identify taxa in both environmental samples and tissue mixtures. The composition and resolution of metabarcoding data depend on the primer(s) used. Markers that amplify different genes can mitigate biases in primer affinity, amplification efficiency, and reference database resolution, but few empirical studies have evaluated markers for complementary performance. Here, we assess the individual and joint performance of 22 markers for detecting species in a DNA pool of >100 species of primarily marine and freshwater fishes, but also including representatives of elasmobranchs, cephalopods, and crustaceans. Marker performance includes the integrated effect of primer specificity and reference availability. We find that a portfolio of four markers targeting 12S, 16S, and multiple regions of COI identifies 100% of reference taxa to family and nearly 60% to species. We then use the four markers in this portfolio to evaluate metabarcoding of heterogeneous tissue mixtures, using experimental fishmeal to test: 1) the tissue input threshold to ensure detection; 2) how read depth scales with tissue abundance; and 3) the effect of non-target material in the mixture on recovery of target taxa. We consistently detect taxa that make up >1% of fishmeal mixtures and can detect taxa at the lowest input level of 0.01%, but rare taxa (<1%) were detected inconsistently across markers and replicates. Read counts showed weak correlation with tissue input, suggesting they are not a valid proxy for relative abundance. Despite this limitation, our results demonstrate the value of a primer portfolio approach—tailored to the taxa of interest—for detecting and identifying both rare and abundant species in heterogeneous tissue mixtures.

Optimizing a metabarcoding primer portfolio for species-level detection of taxa in complex mixtures of diverse fishes

Diana S. Baetscher[1,2*], Nicolas S. Locatelli[1,3], Eugene Won[4], Timothy Fitzgerald[5], Peter B. McIntyre[1], Nina Overgaard Therkildsen[1]

[1]Department of Natural Resources and Environment, Cornell University, Ithaca, NY

[2]Current address: NOAA Alaska Fisheries Science Center, Auke Bay Laboratories, Juneau, AK

[3]Current address: Department of Biology, Penn State, University Park, PA

[4]Department of Animal Science, Cornell University, Ithaca, NY

[5]Environmental Defense Fund, Washington, D.C.

*Corresponding author

Running head

A metabarcoding primer portfolio for fishes

Abstract

DNA metabarcoding is used to enumerate and identify taxa in both environmental samples and tissue mixtures. The composition and resolution of metabarcoding data depend on the primer(s) used. Markers that amplify different genes can mitigate biases in primer affinity, amplification efficiency, and reference database resolution, but few empirical studies have evaluated markers for complementary performance. Here, we assess the individual and joint performance of 22 markers for detecting species in a DNA pool of >100 species of primarily marine and freshwater fishes, but also including representatives of elasmobranchs, cephalopods, and crustaceans. Marker performance includes the integrated effect of primer specificity and reference availability. We find that a portfolio of four markers targeting 12S, 16S, and multiple regions of COI identifies 100% of reference taxa to family and nearly 60% to species. We then use the four markers in this portfolio to evaluate metabarcoding of heterogeneous tissue mixtures, using experimental fishmeal to test: 1) the tissue input threshold to ensure detection; 2) how read depth scales with tissue abundance; and 3) the effect of non-target material in the mixture on recovery of target taxa. We consistently detect taxa that make up >1% of fishmeal mixtures and can detect taxa at the lowest input level of 0.01%, but rare taxa (<1%) were detected inconsistently across markers and replicates. Read counts showed weak correlation with tissue input, suggesting they are not a valid proxy for relative abundance. Despite this limitation, our results demonstrate the value of a primer portfolio approach—tailored to the taxa of interest—for detecting and identifying both rare and abundant species in heterogeneous tissue mixtures.

Keywords

Metabarcoding, primer choice, fishes, tissue mixture, fishmeal

Introduction

Increasingly, ecological studies can leverage DNA metabarcoding to count and identify the species present in an environment or a complex mixture of tissues. Frequent application areas in aquatic ecosystems include plankton surveys, food web analyses, and forensic identification of harvested species. Accurate and effective analysis of the DNA content in these sample types depends on a series of critical methodological decisions, foremost of which is the choice of barcoding primers (Aizpurua et al., 2018; Alberdi et al., 2019; Alberdi, Aizpurua, Gilbert, & Bohmann, 2018; Zhang, Zhao, & Yao, 2020; Zinger et al., 2019). Primer selection influences which taxa can be detected, and the taxonomic resolution to which they can be identified. Barcoding primers amplify sections of genes, which have been selected to provide a balance between having enough divergence to distinguish species and being conservative enough to allow amplification across major taxonomic groups. So-called universal primers, which rely on highly conserved nucleotide binding sites, are attractive because a single marker can amplify a wide range of taxa. However, the greater the breadth of taxa covered (e.g., all metazoa or all teleost fishes), the less likely that species-level identification will be possible because of a lack of sequence resolution when priming sites are conserved across divergent taxonomic groups. Another issue when attempting to obtain species identification is the failure of markers to amplify due to mismatches between primers and template sequences. These mismatches can lead to poor taxon recovery or cause less competitive taxa to drop out if sequencing depth is insufficient (Aizpurua et al., 2018). Primer-template mismatches are more common in diverse samples (Elbrecht et al., 2019); thus, researchers can improve recovery of constituents by combining universal primers that selectively amplify each focal taxonomic group with additional primers that offer species resolution (e.g., Thomsen et al., 2012) or using multiple primers that are optimized for different taxonomic groups (Aizpurua et al., 2018; Berry et al., 2017; Carroll et al., 2019; Evans et al., 2016; Jeunen et al., 2019; Koziol et al., 2019; Silva et al., 2019). Yet, even when using multiple primers, many studies do not obtain species-level assignments because of the challenge of balancing taxonomic breadth and resolution (Djurhuus et al., 2020; Leray & Knowlton, 2015; Locatelli, McIntyre, Therkildsen, & Baetscher, 2020).

Since the DNA in tissue mixtures of interest is often degraded, primers that target short DNA fragments, or minibarcodes, may recover a more complete amplification across taxa in such samples. Smaller barcodes

are more readily amplified than longer fragments and these shorter fragments are more likely to persist in environmental samples (Shokralla et al., 2015; Staats et al., 2016) or stomach contents (Devloo-Delva et al., 2019). Studies that have compared full-length and minibarcodes for mitochondrial Cytochrome c oxidase I (COI) found that minibarcodes 200-300 bp provide comparable resolution to the full-length 658 bp barcode (Hajibabaei et al., 2006; Yeo, Srivathsan, & Meier, 2020). Moreover, full-length barcodes failed to amplify degraded samples (processed fish products), whereas minibarcodes recovered species-level sequences (Marín et al., 2018; Yeo, Srivathsan, & Meier, 2020). Short barcodes are also more economical to sequence than full-length barcoding genes, as current low-cost, high-throughput sequencing platforms tend to produce read lengths of [?] 300 bp. This means that for barcodes shorter than this length researchers can obtain greater read depth for a given investment in sequencing, which can be important because greater sequencing depth potentially detects more rare taxa (Singer, Fahner, Barnes, McCarthy, & Hajibabaei, 2019; Smith & Peay, 2014).

While initial barcoding efforts for animals primarily leveraged variation within the COI gene (e.g., Barcode of Life, Ratnasingham & Hebert, 2007), several other mitochondrial genes have become attractive alternatives (e.g., Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014; Machida & Knowlton, 2012; Miya et al., 2015). The popularity of certain barcoding genes has made extensive high-quality reference data available via the NCBI and BOLD databases to support taxonomic assignments. Availability of suitable reference data for particular taxonomic groups and the accuracy of those data varies among barcoding genes (Leray, Knowlton, Ho, Nguyen, & Machida, 2019), hence it is a key factor in choosing primers.

Aquatic habitats – both marine and freshwater – have become popular targets for metabarcoding studies, likely because of the logistical challenges and considerable expense associated with traditional sampling and survey methodologies (e.g., Salter, Joensen, Kristiansen, Steingrund, & Vestergaard, 2019). A product of these studies are dozens of primer sets for fishes and aquatic taxa which offer researchers an abundance of reference data for interpreting metabarcoding results; yet choosing the optimal primer portfolio also requires assessment of amplification biases and potential sample degradation. To this end, some studies evaluate primers *in silico* and/or in the laboratory, but comparisons have been largely ad hoc and of limited geographic and taxonomic extent. Notably, the results of *in silico* assessments, which frequently guide primer selection, sometimes differ from those of*in vivo* tests (Alberdi et al., 2019; Zhang et al., 2020). The most comprehensive comparison of eDNA and metabarcoding primers for fishes to date (Zhang et al., 2020), for example, assessed primers based exclusively on freshwater fishes from waterbodies in Beijing. Although such an assessment is beneficial, the results may have limited application to marine or endemic species outside this region, and therefore more empirical testing and comparison of the performance and complementarity of metabarcoding markers is needed.

Despite the proliferation of studies using multiple metabarcoding markers, few studies have experimentally tested the additive benefit of a portfolio of markers (each of which amplify a single locus) for obtaining high resolution (species- or genus-level) taxonomic assignments (but see Corse et al., 2019). Instead, many studies that rely on multiple primer sets use each one to identify different taxonomic groups (Berry et al., 2017) or to balance the trade-off between sequence identification at a high taxonomic rank and resolution of taxa within a rank (e.g., Carroll et al., 2019; Djurhuus et al., 2018). However, even within a single taxonomic group, different primers pairs may amplify different subsets of species due to polymorphisms in the primer regions, resulting in complementarity for detection of even closely related taxa. Further complementarity can be gained from varying levels of sequence divergence within the amplified targets, which may result in different markers allowing species-level resolution for different subsets of taxa. Identification to species-level is often important, such as when samples may include closely related species that must be distinguished for biodiversity accounting, fishery and wildlife management, and species conservation. Accordingly, careful design of primer portfolios can boost both the detection rates and resolution of metabarcoding studies, but little empirical testing has explored this potential.

To assess primer complementarity arising from amplification bias, reference data, and trade-offs between taxonomic resolution and breadth, we empirically assess 22 markers, some of which are universal fish primers

and others that are taxon-specific, for their ability to recover species-level identification from a diverse reference DNA pool of >100 species of primarily marine and freshwater fishes, but also including a few representatives of other marine organisms (elasmobranchs, crustaceans, and cephalopods) to evaluate species recovery beyond the target taxonomic group. We then explore the utility of a portfolio approach using complementary markers that amplify sections of COI, 16S, and 12S genes. Marker performance is assessed based on the integrated effect of primer specificity and availability and resolution of reference sequences for the particular taxa in our DNA mixture, and – in this framework – markers are valuable when they contribute species identifications for taxa that are not identified by any other markers. We then test the optimal portfolio from our initial analysis on a set of different tissue mixtures to assess 1) the tissue input threshold to ensure detection; 2) how read depth scales with tissue abundance; and 3) the effect of non-target material in the mixture on recovery of target taxa (marker performance).

Our study was designed to optimize tools for forensic assessment of aquaculture feed composition and accordingly, our DNA pools were composed of aquatic taxa that might be found in fishmeal or other complex tissue mixtures derived from marine and inland fisheries (Mo, Man, & Wong, 2018; Tacon & Metian, 2008) and our tissue mixtures were designed to emulate aquaculture feeds. However, these mock feeds are very similar in nature to other types of tissue mixtures studied broadly in ecology, including stomach contents, fecal samples, and plankton tows. Hence, our overall findings and approach should be transferable to many applications of metabarcoding analysis of heterogeneous tissues.

Methods

Metabarcoding markers

A total of 22 markers for mitochondrial (COI, 12S, 16S) and nuclear (18S, 28S) barcoding genes were identified from metabarcoding, eDNA, and Sanger sequencing barcoding studies that amplified marine and freshwater fishes and tissue mixtures, including seafood products (Table 1). Most of the tested markers were designed to broadly target bony fish (teleosts), but we also included markers specifically targeting elasmobranchs, crustaceans, and cephalopods, taxonomic groups that are often poorly resolved by universal barcodes. We included markers from five different barcoding genes to account for gaps in database reference sequences and limited sequence variation that can lead to poor species-level resolution for certain taxonomic groups in some genes or conversely, too much sequence variation in primer binding sites that can lead to an inability to detect certain taxa (e.g., Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). Only markers that amplified targets <300 bp were selected because minibarcodes are more likely to amplify degraded constituents.

Samples

Reference DNA pools

To compare the amplification performance of all 22 markers, we constructed two pools with equal concentrations of extracted DNA from each of 98 marine and freshwater teleost fish and five representatives of elasmobranch, crustacean, and cephalopod species, in total spanning 88 genera and 60 families (full reference, FR pool; Table 2). Samples were obtained primarily from museums, but also from fish markets when commercially important species were otherwise unavailable. Although we used internal (not surface) tissue for DNA extractions, samples obtained from seafood markets could have come into contact with other fish products, potentially carrying trace contamination that might erroneously be considered false positives if detected by metabarcoding. To better avoid these potential sources of trace contamination, we constructed a second, more restricted reference pool including only the 73 DNA extracts from vouchered museum specimens (vouchered reference, VR).

Experimental tissue mixture samples

Metabarcoding is typically used to detect both rare and abundant constituents in mixtures, and most application cases (diets, larval assemblages, net tows, etc.) include species in unequal proportions along with varying amounts and types of non-target material. To evaluate detection power in this scenario, we used

fishmeal mixed with non-target tissue composed of different fillers. The purpose of the non-target tissue (filler) was to test whether metabarcoding markers are negatively impacted by fillers, either because of a loss of on-target sequencing reads or because of potential PCR inhibition. Our experimental mixtures emulated aquaculture feeds; but the experimental feeds have similarities to other mixed sample types, for example, stomach contents of an omnivore where fish prey items may be mixed with other (non-fish) constituents. In our test case, we freeze-dried muscle tissue from 30 of the unvouchered fish species in the FR pool, coarsely homogenized each sample in a coffee grinder, and then freezer milled samples flash-frozen with liquid nitrogen into a powder. Powdered samples were then assigned to one of six abundance levels and added to make up either 13.33%, 3.65%, 1.91%, 1%, 0.1%, or 0.01% of the total mixture (by weight), thus spanning four orders of magnitude in relative representation from abundant-to-rare (Table 3). Each abundance level was represented by five species. This experimental design allowed us to assess how dominant and rare taxa added at discrete proportions to a heterogenous mixture relate to the proportion of sequencing reads attributed to each taxon and to compare amplification biases across multiple taxa added in the same amount to the fishmeal.

To test how nontarget material or mixture matrix could affect metabarcoding efficiency and therefore, species recovery, the multi-species fishmeal mixtures were combined with two unique filler types to make seven individual experimental feeds with low (2%), medium (10%), and high (25%) ratios of fishmeal-to-filler (Table 3). Fillers for experimental feeds included plant-derived materials – grain and grass flours – and animal byproducts – bloodmeal and feathermeal – to emulate mixture constituents used in fish production (i.e., aquaculture feeds), but are also representative of non-fish diet components. Fishmeal proportions also mimicked potential levels of fish tissue added to aquaculture feeds, from low (0%-2%) to high (25%) proportions of fish in the feed mixture. By multiplying the ratio of fishmeal-to-filler by the fishmeal tissue in the experimental feed, we could test the detection threshold for individual taxa in the overall tissue mixture down to 0.0002% of the experimental feed by mass (i.e., the smallest tissue input – 0.01% of fishmeal – at 2% fishmeal-to-filler).

Laboratory protocol

Eleven of the vouchered samples were obtained as DNA extracts from museums (Table 2). For the remaining samples, DNA was extracted from muscle tissue preserved in ethanol (vouchered specimens), frozen muscle tissue (vouchered specimens and market samples), or dried ground tissue (tissue mixture samples) with single tube column extractions (Omega BioTek EZNA Tissue kits) following the manufacturer's instructions. Extraction blanks were included for each batch of extractions.

DNA extracts were quantified by a Qubit fluorometer (high-sensitivity or broad-range dsDNA assay depending on concentration range), diluted with DNAse-free water, and added in equal proportion to the FR and VR DNA pools. To account for pipetting error, three replicates of the FR and VR DNA pools were constructed by pooling individual DNA extracts in triplicate. For the experimental feeds, to disentangle whether biases in the proportion of sequencing reads attributed to each species were caused by variation in amplification or DNA extraction efficiency, DNA extracts from the 30 fishmeal species were combined in two additional mock DNA pools: one with equal concentration among all taxa (mock equal, ME) and the other in which DNA extract concentration was proportionate to the amount of tissue included in the fishmeal (mock variable, MV). Similar to the previous reference DNA pools, DNA pools for the ME and MV were prepared in triplicate (Fig. S1).

Sequencing library preparation

Metabarcoding sequencing libraries were prepared from each pool using a two-step amplicon protocol (D'Aloia, Bogdanowicz, Harrison, & Buston, 2017) in which an initial 34-cycle PCR targets the gene region of interest using locus-specific primers with Nextera 5' tails (5'-TC GTCGGCAGCGTCAGATGTGTATAAGAGACAG appended to each forward primer and 5' -GTCTCGTGGGCTC GGAGATGTGTATAAGAGACAG to each reverse primer, full reaction details in the Supporting information). Equal volumes of the locus-specific PCR products for each sample were then pooled and a second 5-cycle PCR further amplified

the product and added Nextera-style sequencing adapters with unique i5 and i7 indexes that allow sequencing reads to be assigned to samples during analysis (details about reagent concentrations and PCR conditions in the SI). Rather than using combinatorial indexing, which can lead to mis-assigned reads caused by index-swapping (Caroe & Bohmann, 2020; Schnell, Bohmann, & Gilbert, 2015), we used custom-synthesized adapters with unique dual indexes (Table S1) that can unequivocally identify samples by 8-base indexes on both ends of the molecule. For the initial screening with the FR and VR pools, 16 of the markers were amplified in individual PCRs and six of the markers were duplexed in three reactions – each with two markers. Duplexed markers amplified fragments from different barcoding genes with >75 bp between expected amplicon sizes, which allowed for visualization of two bands, one for each marker, on an agarose gel (see Supporting Information for details). For each sample, PCR products for all markers were pooled into a single indexed library. Three individually indexed PCR replicates were performed for each of the three replicate FR and VR DNA pools, for nine total replicates per DNA pool (the experimental design is illustrated in Fig. S1). The experimental feeds and fishmeal species DNA pools were only analysed with the top four markers (see results), each amplified in a separate PCR. Three individually indexed PCR replicates were performed for each of two DNA extracts from an experimental feed, resulting in six replicates per experimental feed (see SI, Fig. S1 for schematic). One negative control (no template) was included for every 14 PCR reactions and extraction blanks and negative library preparation controls were carried through to sequencing.

Following the indexing PCRs, all libraries were pooled in equal volume, cleaned using 1.8x AMPure XP beads (Beckman Coulter), and then eluted with 50 ul 10 uM Tris-HCl pH 8. A 2% agarose gel was used to confirm that the indexing PCR was successful based on a size shift after the addition of indexes and adapters (˜130 additional bases) when compared to pooled non-indexed samples. The libraries were sequenced using paired-end 150-bp on one lane of a HiSeq X Ten (Novogene, Inc.) with 15% PhiX to account for moderately low library complexity (following Aizpurua et al., 2018).

Bioinformatics

Sequencing reads were assigned to samples using the unique dual indexes and to markers based on the primer sequence. Sequencing adapters and locus-specific markers were removed from paired-end reads using the linked-adapter function in cutadapt (Martin, 2011). Trimmed reads were imported into qiime2 (Bolyen et al., 2019) and data were then denoised, in a process including estimating sequencing error rates from the data, correcting errors, dereplicating sequences, removing chimeras, and then merging overlapping paired-end reads using the dada2 *denoise-paired* function with default parameters (Callahan et al., 2016). Reads were then truncated at +/- 10% of the expected amplicon-size for each locus to remove low quality bases at the trailing ends of sequencing reads (Callahan et al., 2016). The denoising algorithm derives an error model from the sequencing data and aims to correct errors in order to minimize single-nucleotide sequencing errors and therefore, preserve true species-level variation as amplicon sequencing variants (ASV; Callahan, McMurdie, & Holmes, 2017). Because of the short fragment length of the minibarcodes, often just a small number of base-pair differences distinguish species within the same genus (Edgar, 2018), and thus, ASVs have the potential to provide better species-level taxonomic resolution than other analysis methods (e.g., operational taxonomic units, OTU; Callahan, McMurdie, & Holmes, 2017).

Despite similar DNA inputs to PCR reactions, sequencing read depth varied several orders of magnitude among markers for reference DNA pool samples (SI, Fig. S2). To better evaluate marker performance, we subsampled 100,000 reads per marker per sample for each occurrence that included >100,000 reads (two markers, 16Svar and L2513/H2714 had <100,000 reads). Subsampling was done on the fastq files using the *sample* function from seqtk (https://github.com/lh3/seqtk) and then output (subsampled) data were reprocessed in qiime2 and dada2, as described above.

Data for each marker was analysed separately, and the resulting tables included read counts for every ASV sequence occurring in each sample. ASV sequence data were exported from qiime2 in FASTA format and then queried against a custom metazoan database derived from the NCBI Nucleotide sequence database. Separate databases for each gene (12S, 16S, COI, 18S, and 28S) were created by querying NCBI using Entrez. The search was performed using the gene name and a filter that included only animals and sequences

matching the NCBI taxon ID for Metazoa (33208). Gene names were offered to Entrez in many variants (e.g. "COI", "COX1", "cytochrome oxidase subunit I", "cytochrome oxidase subunit 1") to account for variation in naming scheme (additional details and code on GitHub). For mitochondrial genes (12S, 16S, and COI), queries were also restricted to mitochondrial sequences and filtered by length to exclude sequences substantially shorter than any of the targeted amplicons (<50 bp) while including whole mitochondria of all sizes (filter <100,000 bp; largest known metazoan mitochondrial genome is 80,923 bp; Stampar et al., 2019). For nuclear genes (18S and 28S), the same length filters were used to exclude short sequences (<50 bp) and long genomic scaffolds (<100,000 bp). Search terms "scaffold" and "shotgun" were excluded from search queries. These filtered Entrez queries were then downloaded separately in FASTA format. By using only NCBI nucleotide data for genes that corresponded to our markers, and only searching the custom database that corresponded to the appropriate gene for each primer set, we were able to reduce overall computation time and potential off-target BLAST hits. Downloaded sequences were further filtered to remove environmental, unverified, uncultured, and protein database sequences and then configured into a BLAST database with makeblastdb from the BLAST+ suite (Camacho et al., 2009; code available on GitHub).

FASTA output from qiime2 was queried against the reference databases using BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) command line with the blastn program using 96% minimum identity, 95% query coverage; e-value = 10, reward = 2 and penalty = -3, penalty to open a gap = 5 and to extend a gap = 2, and with no limit to the number of hits per sequence (no culling limit). The BLAST query might have included some liberal matches, but all BLAST hits were further filtered and refined in subsequent steps (the highest e-value for data passing filter = 1.54 e^{-38}). GenBank accession numbers were used to access the taxonomic lineage for each hit using a custom bash script which utilizes NCBI's accession2taxid files and TaxonKit (Shen & Xiong, 2019) to extract seven taxonomic ranks (i.e., domain, phylum, class, order, family, genus, species) for each BLAST hit (code on GitHub).

GNU Parallel was used throughout the bioinformatic pipeline to run multiple jobs simultaneously while maintaining the optimal number of threads for each program (Tange, 2011).

Data decontamination

Before assigning taxonomic information to each ASV and collectively to every sample, sequence data were filtered to reduce the influence of contamination and errors. First, read count tables were filtered to exclude ASVs without a taxonomic match in the custom BLAST database. This removed non-metazoan sequences and those without confident taxonomic matches (<96% sequence identity in the top hit). Since we sequenced samples deeply and using a two-step PCR protocol, which can amplify errors and contaminants, species occupancy detection modeling (SODM) is an appropriate alternative to using a minimum read depth threshold for handling potential false detections (Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016). We used a Bayesian SODM (Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016) in R to estimate the probability of ASV occurrence for each sample where each of the 6-9 technical replicates for that sample is treated as an independent draw from a binomial distribution (true positives and false positives). Any ASV with an estimated probability of occurrence <80% was removed from the dataset (as in Djurhuus et al., 2020).

Another way to identify potential contamination is by comparing the sequence data among extraction and PCR replicates for the same sample. When replicates are more dissimilar than concordant, this can indicate anomalies during laboratory activities. To compare the sequence composition across replicates of each locus for every sample, the ASV read count data that remained following SODM were analyzed for dissimilarities based on Bray-Curtis distance and NMDS in the R package VEGAN (Oksanen et al., 2019). Sample replicates with Bray-Curtis dissimilarity >0.49 were removed (as in Djurhuus et al. 2020; SI Table S2) and three markers (Teleo, Crust2, and 16Sfish) with insufficient data at this stage were dropped from further analyses; however, these markers could have potentially provided useful data had they received sufficient read depth.

Taxonomic assignment

Separate from the sample-based data decontamination procedure, described above, taxonomic assignment for each metabarcoding sequence required evaluating the full set of BLAST hits for each ASV using a custom R script (R Core Team, 2019). The goal of the R script was to obtain the highest taxonomic resolution for each sequence while accounting for all BLAST hits above the 96% minimum identity required by the blastn query. Species-level identification was only accepted if the ASV sequence matched the database reference sequence at >98% identity (as in Alberdi et al., 2018) and only then if no BLAST hits within 2% identity of the top hit matched a different species. When BLAST results for a given ASV violated either of these rules, the next taxonomic level (i.e., common ancestor) was tested using the same criteria and so on until a consensus taxonomic rank was obtained within the top 2% identity of matches. For example, when an ASV had only a significant hit to a single species, that species was assigned unless the sequence match was <98% identity, in which case, the ASV would be assigned to the genus-level. However, when an ASV had significant hits to multiple taxa, the common ancestor for BLAST hits within the top 2% identity determined whether that sequence could be attributed to a species, genus, or family, or whether the sequence provided little informative variation for a high-resolution assignment (code available on GitHub).

Decontaminated ASV and read count data were merged with taxonomic information from ranked and filtered BLAST hits. Multiple ASVs within a locus that matched the same taxon at more than one taxonomic level (e.g., one ASV identifies the family Clupeidae and another matches the genus *Clupea*) were merged to retain the highest-resolution assignment (in this example, the genus, *Clupea*) for each taxon within each replicate/locus. We reasoned that both sequences would likely come from the same fish, and therefore retained the higher-resolution assignment.

Finally, taxonomic assignments were used to compare the performance of the individual markers and metabarcoding loci for recovering species added to the vouchered reference and full reference DNA pools, as well as determining the optimal combination of markers to maximize identification of reference taxa to species-level.

The portfolio of complementary markers was identified by ranking markers using an accumulation curve to identify which recovered the greatest number of species from the FR, followed by the greatest number of additional species, and so on until the curve saturated (Fig. 1). The minimal panel of primer pairs that captured the full species diversity in the DNA pools were used to analyse the experimental feeds and examine quantitative relationships between relative tissue abundance and sequencing read proportions in heterogeneous mixtures.

Results

Sequencing results

A total of 277,021,343 merged paired-end sequencing reads passed the initial bioinformatic filtering (˜88% of the total raw reads). From these filtered reads, we identified 62,764 unique ASVs. Only 12,792 of those ASVs matched taxa in the database, but this set of ASVs with taxonomic matches accounted for 88.2% of the reads. Accordingly, non-metazoan ASVs (those without a taxonomic assignment) accounted for only ˜12% of the reads.

Libraries for blanks and negatives yielded an order of magnitude fewer reads on average than the samples (SI, Fig. S3), and 30% of those reads and 85% of the associated ASVs did not match taxa in the metazoan database. In total, 3.5% of the metazoan ASVs detected across all samples were only found in the blanks and negative control samples. To eliminate the potential influence of such DNA contamination, some studies have removed all taxa detected in negative controls from their results. Although there is appeal in taking a more quantitative approach and subtracting the read counts for contaminant taxa in blanks from read counts across all samples, this subtraction approach does not account for the lack of template DNA in blanks and negative controls, which leads any DNA contamination to be heavily amplified (McKnight et al., 2019). Instead of subtracting read counts, we attempted to use the*microDecon* R package (McKnight et al., 2019) to subtract reads proportionately for taxa found in negative controls; however, this had the effect of entirely eliminating species that were truly present in our reference DNA pool—a type-II error. To avoid introducing such false negatives, we reported (SI, Fig. S4) – instead of subtracted – these reads for our samples.

Over 55% of the 12,792 ASVs that provided taxonomic information based on our metazoan reference database matched database sequences at the level of species or genus according to our assignment rule (see above). For ASVs that matched at least one taxon in our database, the number of significant BLAST hits ranged from 1 to 500 taxonomic hits (mean = 123; median = 34) and 23.1% of these ASVs recovered species-level assignments (2,956 ASVs), 32.7% of ASVs were assigned to genus (4,181 ASVs), 13.2% to family (1,685 ASVs), 7.7% to order (986 ASVs), 13.8% to class (1,769 ASVs), 0.6% to phylum (76 ASVs), and 8.9% to domain (1,139 ASVs).

Detection success and multiple primer set complementarity

To allow fair comparison, we subsampled reads to improve evenness of coverage across markers and removed three markers with insufficient data (<1,000 reads or <0.1% of the mean number of reads per locus) from further analyses (Teleo, Crust2, and 16Sfish). One additional primer, 16Svar, had low yield (<10,000 reads, <1% of the mean), but included sufficient data for data decontamination so was retained for analysis. Out of the 19 retained markers, a combination of four identified all 60 families of marine and freshwater taxa in the full reference DNA pool. These four markers, FishminiA, nsCOIFo, MiFish, and CEP (for details, see Table 1) provided sufficient taxonomic resolution to correctly recover the genus of 90.9% of taxa and identify to species 58.6% of input taxa (Fig. 1). All but one of the of the species in our reference pool (*Petrochromis kazumbe*) had reference data for at least one of the four markers, so the frequent lack of species-level detection resulted from insufficient sequence variation within the amplified target rather than database representation. Two additional markers (aquaF2 and either aquaF3 or shark474) allowed recovery to genus-level of three more reference taxa (83 of 88 genera; 94.3%) and adding two additional markers (aquaF2 and Fish_COILBC) identified two of the remaining known taxa to species level, but the remaining 13 markers did not. Genus-level assignments were more successful than species-level assignment because BLAST hits to multiple unique species within the top 2% of hits were aggregated to the genus-level.

Marker performance was broadly consistent across taxonomic levels (species, genus, family), with COI markers generally performing better than other barcoding genes. This success was at least partially attributable to the more extensive coverage of our focal taxa in the reference database for COI (SI, Fig. S5). The two top performing markers target adjacent but non-overlapping regions of the COI gene, and of these, the single best marker identified 90% of reference taxa to family, 78.4% to genus, and 41.7% to species (Fig. 2). In combination, these two COI markers identified 95% of families, 85% of them to genus level, and just under 50% to species-level. The best 16S and 12S markers recovered fewer taxa to species-level (˜25% each), but contributed taxa not identified by any other primer, supporting the value of the portfolio (Fig. 2).

For taxon-specific markers, the COI markers for elasmobranchs and plankton identified nearly as many reference taxa (the majority of which were teleosts) as the top-performing COI marker, and with similar taxonomic resolution, and thus were of more general use. However, crustacean and cephalopod markers had limited use outside of these targeted groups (Fig. 2). In contrast, the more general fish markers successfully identified the few representative elasmobranch, crustacean, and cephalopod samples included in our DNA pool, suggesting broader taxonomic reach of those primers.

False positives in the vouchered reference DNA pool

Although the FR community included specimens widely sourced to maximize taxonomic diversity, some of these samples were almost certainly exposed to unknown fish and other metazoan contaminants throughout the supply chain (e.g., market samples). Therefore, we used the vouchered samples and the VR DNA pool to test data decontamination and quantify false positive detections (i.e., detection of taxa not added to our pool).

For all markers, the analysis returned several false positives for the VR pool libraries. The species occupancy modeling and Bray-Curtis data decontamination procedure removed some of these ASVs that corresponded to species known to be present in the laboratory where DNA was extracted, and libraries prepared (Fig. 3). Certain additional taxa may have accompanied vouchered specimens through logical mechanisms, such as parasitic organisms (e.g., *Kudoa sp* .). Species occupancy detection modeling (SODM) removed 18 ASVs

with low estimated probability of occupancy, 11 of which corresponded to species-level taxonomy. Another five false positives were filtered by removing sample replicates with Bray-Curtis dissimilarity scores >0.49 (Fig. 3; SI Table S2). Remaining false positives ranged from 0-7 per locus (mean = 2.9) and included taxa known to be present in the lab, including species added to the FR DNA pool. When tallying only unknown contaminant sources, the number of false positives was reduced to 0-6 per locus (mean = 2.4; Fig. 3).

Experimental tissue mixture samples

Across the portfolio of four top-performing markers, sequencing read depth did not provide a consistent quantitative proxy for the relative amount of tissue added to the fishmeal mixture. While there was a tendency for species with lower input amounts to receive fewer reads, the proportion of reads varied substantially among species that were added in the same amounts (Fig. 4, Fig. 6A). The mean proportion of reads for the five taxa added in the same amount showed a positive relationship for tissue inputs <1%, with the lowest tissue input (0.01%) receiving an order-of-magnitude smaller proportion of reads across all four markers (1.7×10^{-4} of reads) than the next lowest tissue input (0.1%; $5.3 \times 10^{-3}$of reads; SI, Fig. S6). However, the relationship between read proportion and tissue input was inconsistent because the 1.91% tissue input level received, on average, a smaller proportion of reads than taxa that comprised 1.0% of the tissue input (5.6×10^{-3} and 2.7×10^{-2} of reads, respectively) and for three of the four markers, the highest tissue input level (13.32% of fishmeal) received a smaller proportion of reads than taxa comprising 3.65% of the fishmeal mixture (Fig. 4 and SI, Fig. S6).

The ratio of fishmeal-to-filler in the experimental feeds had little effect on the proportion of reads per taxon (Fig. 4). Taxa that comprised 0.01% and 0.1% of the fishmeal showed more variation than taxa present in higher proportions, with inconsistent read depth across fishmeal-to-filler ratios (2%, 10% or 25% fishmeal) and across the four markers (Fig. 4). Although read depths were inconsistent in terms of presence/absence detection, fewer of the low-input taxa were recovered, and this pattern of taxon drop-out increased as the ratio of fishmeal-to-filler decreased (SI, Fig. S7).

Matrix composition (filler) of experimental feeds did not impact the recovery of reference taxa or the proportion of reads attributed to those taxa for constituents that make up >1% of fishmeal. Patterns in proportion of reads were dominated by variation among taxa rather than between filler types, and the proportion of reads did not change between the soy filler and the animal/plant filler for taxa added at >1% fishmeal (Fig. 5). However, low-input taxa, those that comprised 0.01% or 0.1% of fishmeal, displayed more variable read depth between fillers, but not in a consistent way within or across markers. At the lowest tissue input (0.01%), two markers (12S; MiFish and one COI primer set; nsCOIFo) consistently showed a higher proportion of taxon drop-out (SI, Fig. S7).

Although two feeds consisted entirely of filler with no added fishmeal, sequencing results showed that 80% of fishmeal taxa were present in these filler-only feeds, indicating that simultaneous preparation of filler-only feeds and fishmeal feeds resulted in contamination that was sufficient to be detected by our PCR assay (SI, Fig. S8). Sequencing read counts for filler 1 (soy flour) were lower than for filler 2, which contained bloodmeal (from *Sus sp.*) and feathermeal (from*Gallus sp.*). Both taxa were detected by all four of the markers designed to amplify fishes, and in the absence of added fishmeal tissue,*Sus sp.* and *Gallus sp.* received an average of ˜25% of the sequencing reads per replicate, whereas in the experimental feeds that contained 75% filler 2 and 25% fishmeal,*Sus sp.* and *Gallus sp.* received only ˜1% of sequencing reads, highlighting the influence of template competition during library preparation.

Comparing taxon recovery from the 100% fishmeal and pooled DNA extracts (MV) showed that reference taxa were detected (presence/absence) equally well regardless of relative representation when tissue input/DNA concentration was >1% (Fig. 6A, 6B). For input categories below 1% (0.01% and 0.1%), more taxa were identified from the 100% fishmeal sample than from the MV (that contained DNA extracted separately from each experimental feed fish constituent and pooled in proportions equivalent to relative tissue inputs to the feeds). In comparison, when the DNA extracts from each experimental feed constituent was pooled in equal concentration (the ME pool), we recovered all taxa with at least one locus with just one

exception (*Ophisternon sp.*), indicating that taxon drop-out in the MV pool was not a result of primer mismatch but rather a result of low input concentration (all species except for *Loliolus sp.* were consistently identified at concentrations above 2.24 x 10^{-4} ng/ul; Fig. 6B, 6C).

Discussion

We found that a portfolio of four markers targeting three different barcoding genes identified as many of our set of 103 marine and freshwater taxa as possible to the species level and that no additional taxa could be identified by applying an additional 15 markers to our DNA reference pool. These same four markers identified the full set of 103 taxa to the family level. COI markers recovered the greatest number of reference taxa and at the highest taxonomic resolution, but markers for 12S and 16S identified unique taxa missed by COI. These patterns offer strong evidence of the benefits of a portfolio approach, but also suggest that primer portfolios may need to be optimized for each target taxonomic group or geographic region rather than considered broadly transferable.

While we found an overall tendency towards lower read counts for taxa added in lower input amounts, the patterns were highly inconsistent. Accordingly, our read counts could not reliably be interpreted as even a semi-quantitative proxy for relative abundance of a species in our experimental tissue mixture, highlighting a key limitation for PCR-based assays. Detection (presence/absence) of reference taxa from this complex mixture became inconsistent with <1% of input tissue, which points to a threshold of ˜1% representation for robust PCR-based detection in a heterogeneous sample. Interestingly, the likelihoods of detecting species across the range of relative abundance in the mixture was independent of whether fish tissues were analyzed alone or after being diluted with either a plant-based or plant/animal filler, probably thanks to the relative affinity of the markers tested for fishes.

Locus complementarity and taxonomic resolution

Species-level identification in the DNA pools proved more challenging than either genus or family identification, even though all taxa were represented by one or more of the reference databases for the target genes (Fig. S4). Species resolution assignment is more desirable, but also more challenging to obtain because barcoding genes often include insufficient variation to confidently distinguish congeneric species. The challenge of amplifying taxonomically informative variation is particularly true for minibarcodes, which capture a smaller section of larger barcoding genes. Furthermore, reference database information (i.e., GenBank, BOLD) is less complete at the species-level than for genera or families (i.e. some databases lack data for individual species, but a much higher proportion of genera and families are represented) and databases are less accurate for species- and genus-level identification (Leray et al., 2019; Locatelli et al., 2020). For these reasons, biodiversity studies may choose to assign data to family, class, or order, rather than species in order to capture greater taxonomic breadth (e.g., Djurhuus et al., 2020; Leray & Knowlton, 2015). Our results affirm that such an approach would accurately detect 100% of families present in our reference DNA pool.

Notably, two of the top-performing markers amplified adjacent, non-overlapping regions of the COI gene. COI markers benefit from the most complete reference database of the genes we tested (SI, Fig. S4), which is consistent with prior studies of fish tissues (Devloo-Delva et al., 2019). The strategy of including multiple markers for the same gene has been applied often in plant barcoding as well as for 18S rRNA in animals (e.g., Coghlan, Shafer, & Freeland, 2020; Machida & Knowlton, 2012). Fewer studies show the added benefit of multiple COI markers (but see Corse et al., 2019; Shokralla, Hellberg, Handy, King, & Hajibabaei, 2015; Valdez-Moreno et al., 2019). Including two COI minibarcode markers in our portfolio hedges against the limitations of amplifying degraded samples while leveraging the robust COI reference data for diverse marine and freshwater taxa.

Despite the popularity of 12S for metabarcoding marine and freshwater fishes, and the commensurate abundance of reference data (Miya et al., 2015; Masaki Miya, Gotoh, & Sado, 2020), our top 12S primer set identified fewer reference taxa than the top COI and 16S markers. However, the 12S locus contributed more unique species-level identifications that were not recovered by other genes (Fig. 2), hinting at the overall utility of this region for barcoding fish to the species level. Coupled with results from Zhang et al. (2020), in

which 12S markers identified the largest number of fishes from waterbodies in Beijing, our results reinforce the expectation that optimal markers may differ across habitats and taxonomic groups, even within fishes.

Markers targeting specific taxonomic groups – sharks, plankton, crustaceans, and cephalopods – provided no additional resolution for reference taxa in the DNA pools (because our representatives from these taxa were detected with our top-performing teleost fish primers). Surprisingly, COI markers designed for sharks and plankton performed nearly as well on teleost fish as the best universal fish COI markers. However, the opposite was true for crustacean and cephalopod markers, which had little utility outside their targeted taxonomic groups. Admittedly, we had few representatives of these groups in the DNA pools to test the potential increased resolution of taxon-specific markers, so our results are not conclusive, but suggest that markers can show high performance outside their immediate target group (Fig. 2, 3).

Both 18S markers included a higher proportion of false positives and contamination in the extraction blanks and PCR negatives than other gene regions, possibly due to a mismatch between the resolution of the 18S barcoding region and the species composition of the DNA pools (e.g., 18S may be better for identifying diverse groups to class or order and consequently picks up more bacterial contamination; SI, Fig. S4). A similar explanation – non-specific amplification – may account for the limited number of target taxa amplified by the lone 28S locus. Interestingly, a prior study noted that non-specific amplification in COI markers impaired eDNA analyses for marine and freshwater fishes (Collins et al., 2019); yet this study did not test either of our top-performing COI markers, illustrating both the impressive number of universal fish COI markers and that non-specific amplification resulting in false detections can vary among markers within a single barcoding gene and for different applications, i.e., tissue mixture metabarcoding or eDNA.

Unfortunately, three markers that have amplified well in other studies (e.g., Polanco et al., 2021; Pont et al., 2018) got so few sequencing reads that we were unable to retain them in our analysis. The three markers that dropped out were also those that, based on preliminary data (agarose gel bands), we chose to amplify in multiplex PCR reactions (paired with one additional marker, in each case). However, for each of the three multiplexes, one marker performed well, and one did not. Thus, our exploration of multiplex reactions revealed challenges that would require taking amplified products through to sequencing in order to confirm that both markers receive a comparable number of reads (De Barba et al., 2014). Despite the validation steps necessary for effective multiplexing, doing so with complementary markers that amplify different barcoding genes could ultimately yield a more efficient laboratory workflow.

Taken together, our results underscore the advantages of using an optimized portfolio of barcoding markers (similar to results described by Shaw et al., 2016; Zhang, Zhao, & Yao, 2020), yet also reveal that adding markers to a portfolio without testing for complementarity can increase project costs and laboratory effort without improving detection or identification. Further, additional markers can increase the number of false positive observations – by nontarget amplification or mismatches with reference data – and these issues can be more acute when researchers seek high-resolution species identification from broad biodiversity surveys. For studies aiming to quantify biodiversity based on sequence variation patterns, researchers should also be aware of potential nontarget amplification of nuclear mitochondrial pseudo-genes (numts), and can use available software (i.e., *metaMATE* , Andujar et al., 2021) to remove these sources of error.

Experimental tissue mixture samples

To explore how a metabarcoding primer portfolio performs on heterogeneous tissue mixtures with varying amounts of each constituent, such as for applications to aquaculture ingredient tracing, we created a complex fishmeal mixture that was further diluted with fillers. Across these experimental feeds, the proportion of sequencing reads recovered did not reliably reflect relative amounts of input tissue quantity, although taxa input at <1% of the fishmeal mixture consistently received a smaller proportion of reads and went undetected by markers more frequently than taxa that comprised at least 1% of the fishmeal. Squid (*Loliolus sp.*) was the exception, likely because of inefficient DNA extractions from cephalopod tissues using traditional methods (Fig. 6; Lee, McFall-Ngai, Callaerts, & de Couet, 2009). Accordingly, our results are in line with conclusions from a recent review that suggested that a weak quantitative relationship may exist between

relative DNA input amounts and sequence yields, albeit with a large degree of uncertainty (Lamb et al., 2019). Ecological studies of tissue mixtures or diets are typically not accompanied by sensitivity analyses of metabarcoding assays to assess detection limits, but food science applications using untargeted deep sequencing of genomic DNA also identified mixture constituents at >1% of an experimental composition (Haiminen et al., 2019; Ripp et al., 2014), concordant with our PCR-based results.

Poor recovery of *Loliolus sp.* highlights a key observation that although primer-binding and PCR amplification biases receive considerable attention in the metabarcoding literature (e.g. Deagle et al., 2019; Elbrecht & Leese, 2015), variation in DNA extraction efficiency from different tissue types and taxa may, in some cases, be the factor that undermines quantitative inferences from sequencing reads. One line of evidence for tissue extraction bias in our study comes from the consistent performance of certain taxa added to the fishmeal mixture across four markers that amplify three different barcoding genes (COI, 12S, 16S, Fig. 6). If PCR bias were the dominant effect, presumably the amplification bias would favor different taxa for each primer set. Instead, we conclude that discrepancies in DNA extraction efficiency are the most likely explanation for our result that *Scomber scombrus* and *Salmo salar* acquire the largest share of sequencing reads across all three genes (four markers).

Variation in read counts among the five species added in equal mass to our fish mixture suggests that idiosyncrasies of particular taxa are at least as important as the actual tissue input amount. Variation among taxa in low-input categories (0.01% and 0.1%) also could arise from heterogeneity within our fishmeal mixture. Inconsistent read proportions per taxon among markers, among fishmeal percentages in experimental feeds (Fig. 4), and between filler types (Fig. 5), as well as taxon drop-out (Fig. 6), highlight the uncertainty associated with accurately recovering constituents that comprise the smallest proportions of a mixture. Further, PCR repeatability becomes less reliable for very low DNA inputs (SI, Fig. S7). Although metabarcoding studies often filter data based on taxon occurrence in multiple technical replicates (Alberdi et al., 2018), this creates a conservative bias with respect to truly rare species.

The filler composition of experimental feed samples did not impact the recovery of reference taxa or proportion of sequencing reads attributed to those taxa, likely because the filler was sufficiently taxonomically divergent that universal fish (teleost) primers preferentially amplified the target DNA. Yet if filler competes with target DNA during PCR amplification – either because of lack of locus-specificity or taxonomic similarity between target taxa and matrix – then shallow sequencing depth could affect recovery of target taxa, especially those added in very small amounts. Here, we recovered most of the lowest tissue-input taxa (0.01%) with one or more markers, suggesting that sequencing depth is not a limiting factor in the present study.

Although PCR-free and untargeted approaches appear better-suited to quantitative inference of relative representation (Haiminen et al., 2019), these also require more starting material (Haynes, Jimenez, Pardo, & Helyar, 2019) and genome reference data, which is currently unavailable for many of the 30 reference taxa added to our experimental feeds. Further, the challenges of tissue mixtures – low inputs of fragmented DNA and the presence of inhibitors – will likely remain problematic for accurate quantification (Haynes et al., 2019), particularly as mixture complexity and heterogeneity increases.

Conclusion

Our study demonstrates that a portfolio of metabarcoding markers provides sufficient taxonomic coverage to detect and identify all of the >100 marine and freshwater taxa collected worldwide and combined here into heterogeneous and complex mixtures. Although multi-locus metabarcoding has become increasingly common for biodiversity assessments, where studies rely on different loci (COI, 18S, 16S, 12S) to resolve different taxonomic groups, our focus was to employ multiple markers to provide greater taxonomic coverage and increased taxonomic resolution across a single group: fishes (and tangentially, other aquatic taxa). We recommend that investigators seek to optimize their own portfolios of markers suited to the taxonomic and geographic scope of their work, but our findings demonstrate that this approach can enable powerful inferences about biodiversity. Unfortunately, in order to achieve quantitative inferences about relative

abundance of each species in a complex mixture, we must recommend further exploration of capture-based approaches (e.g., Mariac et al., 2018) or leveraging genome-wide data (e.g., skim-seq; Bohmann, Mirarab, Bafna, & Gilbert, 2020; Chua et al., 2021; Kobus et al., 2020). Accurately detecting and then quantitatively accounting for the relative abundance species in complex mixtures is a critical next step for ecological and forensic studies using eDNA.

Acknowledgements

For vouchered specimens, we thank Heath Cook and Willy Bemis and the Cornell University Museum of Vertebrates, Kansas University, the Northeastern University Ocean Genome Legacy Center, and the Smithsonian National Museum of Natural History collections. Tissues in our feed mixture were derived from collections by P.B.M. under permission from the nations of Venezuela and Tanzania, and the State of New York. Thanks to Harmony Borchardt-Wier for laboratory assistance. This work was supported by a gift from the David R. and Patricia D. Atkinson Foundation to the Cornell Atkinson Center for Sustainability and Environmental Defense Fund to P.B.M. and N.O.T.

References

Aizpurua, O., Budinski, I., Georgiakakis, P., Gopalakrishnan, S., Ibanez, C., Mata, V., ... Alberdi, A. (2018). Agriculture shapes the trophic niche of a bat preying on multiple pest arthropods across Europe: Evidence from DNA metabarcoding. *Molecular Ecology* ,*27* (3), 815–825. doi: 10.1111/mec.14474

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., & Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources* , *19* (2), 327–348. doi: 10.1111/1755-0998.12960

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution* , *9* (1), 134–147. doi: 10.1111/2041-210X.12849

Altschul, Stephen F., Gish, Warren, Miller, W., Myers, Eugene W., & Lipman, David J. (1990). Basic Local Alignment Search Tool.*Journal of Molecular Biology* , *215* (3), 403–410.

Berry, T. E., Osterrieder, S. K., Murray, D. C., Coghlan, M. L., Richardson, A. J., Grealy, A. K., ... Bunce, M. (2017). DNA metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (Neophoca cinerea). *Ecology and Evolution* , *7* (14), 5435–5453. doi: 10.1002/ece3.3123

Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology* , *29* (14), 2521–2534. doi: 10.1111/mec.15507

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J.-B., ... Mouillot, D. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances* ,*4* (5), eaap9661. doi: 10.1126/sciadv.aap9661

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* , *11* (12), 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* , *13* (7), 581–583. doi: 10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* , *10* (1), 421. doi: 10.1186/1471-2105-10-421

Cardenosa, D., Fields, A., Abercrombie, D., Feldheim, K., Shea, S. K. H., & Chapman, D. D. (2017). A multiplex PCR mini-barcode assay to identify processed shark products in the global trade. *PLOS ONE* ,*12*

(10), e0185368. doi: 10.1371/journal.pone.0185368

Caroe, C., & Bohmann, K. (2020). *Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples* [Preprint]. Molecular Biology. doi: 10.1101/2020.01.22.915009

Carroll, E. L., Gallego, R., Sewell, M. A., Zeldis, J., Ranjard, L., Ross, H. A., ... Constantine, R. (2019). Multi-locus DNA metabarcoding of zooplankton communities and scat reveal trophic interactions of a generalist predator. *Scientific Reports* ,*9* (1), 281. doi: 10.1038/s41598-018-36478-x

Chua, P. Y. S., Crampton-Platt, A., Lammers, Y., Alsos, I. G., Boessenkool, S., & Bohmann, K. (2021). Metagenomics: A viable tool for reconstructing herbivore diet. *Molecular Ecology Resources* ,*21* (7), 2249–2263. doi: 10.1111/1755-0998.13425

Coghlan, S. A., Shafer, A. B. A., & Freeland, J. R. (2020). Development of an environmental DNA metabarcoding assay for aquatic vascular plant communities. *Environmental DNA* , edn3.120. doi: 10.1002/edn3.120

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., ... Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution* , *10* (11), 1985–2001. doi: 10.1111/2041-210X.13276

Corse, E., Tougard, C., Archambaud-Suard, G., Agnese, J., Messu Mandeng, F. D., Bilong Bilong, C. F., ... Dubut, V. (2019). One-locus-several-primers: A strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding studies. *Ecology and Evolution* , *9* (8), 4603–4620. doi: 10.1002/ece3.5063

D'Aloia, C. C., Bogdanowicz, S. M., Harrison, R. G., & Buston, P. M. (2017). Cryptic genetic diversity and spatial patterns of admixture within Belizean marine reserves. *Conservation Genetics* ,*18* (1), 211–223. doi: 10.1007/s10592-016-0895-5

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources* , *14* (2), 306–323. doi: 10.1111/1755-0998.12188

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome oxidase subunit I marker: Not a perfect match. *Biology Letters* , *10* (9), 20140562. doi: 10.1098/rsbl.2014.0562

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., ... Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology* , *28* (2), 391–406. doi: 10.1111/mec.14734

Devloo-Delva, F., Huerlimann, R., Chua, G., Matley, J. K., Heupel, M. R., Simpfendorfer, C. A., & Maes, G. E. (2019). How does marker choice affect your diet analysis: Comparing genetic markers and digestion levels for diet metabarcoding of tropical-reef piscivores. *Marine and Freshwater Research* , *70* (1), 8. doi: 10.1071/MF17209

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., ... Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications* , *11* (1), 254. doi: 10.1038/s41467-019-14105-1

Djurhuus, A., Pitz, K., Sawaya, N. A., Rojas-Marquez, J., Michaud, B., Montes, E., ... Breitbart, M. (2018). Evaluation of marine zooplankton community structure through environmental DNA metabarcoding: Metabarcoding zooplankton from eDNA. *Limnology and Oceanography: Methods* , *16* (4), 209–221. doi: 10.1002/lom3.10237

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* , *34* (14), 2371–2375. doi: 10.1093/bioinformatics/bty113

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., ... Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* ,*7* , e7745.

doi: 10.7717/peerj.7745

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLOS ONE* , *10* (7), e0130324. doi: 10.1371/journal.pone.0130324

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., ... Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources* , *16* (1), 29–41. doi: 10.1111/1755-0998.12433

Fernandes, T. J. R., Silva, C. R., Costa, J., Oliveira, M. B. P. P., & Mafra, I. (2017). High resolution melting analysis of a COI mini-barcode as a new approach for Penaeidae shrimp species discrimination. *Food Control* , *82* , 8–17. doi: 10.1016/j.foodcont.2017.06.016

Fields, A. T., Abercrombie, D. L., Eng, R., Feldheim, K., & Chapman, D. D. (2015). A Novel Mini-DNA Barcoding Assay to Identify Processed Fins from Internationally Protected Shark Species. *PLOS ONE* ,*10* (2), e0114844. doi: 10.1371/journal.pone.0114844

Gunther, B., Knebelsberger, T., Neumann, H., Laakmann, S., & Martinez Arbizu, P. (2018). Metabarcoding of marine environmental DNA based on mitochondrial and nuclear genes. *Scientific Reports* , *8* (1), 14822. doi: 10.1038/s41598-018-32917-x

Haiminen, N., Edlund, S., Chambliss, D., Kunitomi, M., Weimer, B. C., Ganesan, B., ... Beck, K. L. (2019). Food authentication from shotgun sequencing reads with an application on high protein powders. *Npj Science of Food* , *3* (1), 24. doi: 10.1038/s41538-019-0056-6

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded: BARCODING. *Molecular Ecology Notes* , *6* (4), 959–964. doi: 10.1111/j.1471-8286.2006.01470.x

Haynes, E., Jimenez, E., Pardo, M. A., & Helyar, S. J. (2019). The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. *Food Control* , *101* , 134–143. doi: 10.1016/j.foodcont.2019.02.010

Horreo, J. L., Ardura, A., Pola, I. G., Martinez, J. L., & Garcia-Vazquez, E. (2013). Universal primers for species authentication of animal foodstuff in a single polymerase chain reaction: Universal primers for species authentication. *Journal of the Science of Food and Agriculture* , *93* (2), 354–361. doi: 10.1002/jsfa.5766

Jeunen, G., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., ... Gemmell, N. J. (2019). Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. *Molecular Ecology Resources* , *19* (2), 426–438. doi: 10.1111/1755-0998.12982

Kitano, T., Umetsu, K., Tian, W., & Osawa, M. (2007). Two universal primer sets for species identification among vertebrates. *International Journal of Legal Medicine* , *121* (5), 423–427. doi: 10.1007/s00414-006-0113-y

Kobus, R., Abuin, J. M., Muller, A., Hellmann, S. L., Pichel, J. C., Pena, T. F., ... Schmidt, B. (2020). A big data approach to metagenomics for all-food-sequencing. *BMC Bioinformatics* ,*21* (1), 102. doi: 10.1186/s12859-020-3429-6

Koziol, A., Stat, M., Simpson, T., Jarman, S., DiBattista, J. D., Harvey, E. S., ... Bunce, M. (2019). Environmental DNA metabarcoding studies are critically affected by substrate selection. *Molecular Ecology Resources* , *19* (2), 366–376. doi: 10.1111/1755-0998.12971

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources* ,*16* (3), 673–685. doi: 10.1111/1755-0998.12486

Lee, P. N., McFall-Ngai, M. J., Callaerts, P., & de Couet, H. G. (2009). Preparation of genomic DNA from Hawaiian bobtail squid (Euprymna scolopes) tissue by cesium chloride gradient centrifugation. *Cold Spring*

Harbor Protocols , *2009* (11), doi: 10.1101/pdb.prot5319

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity.*Proceedings of the National Academy of Sciences* , *112* (7), 2076–2081. doi: 10.1073/pnas.1424997112

Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences* ,*116* (45), 22651–22656. doi: 10.1073/pnas.1911714116

Locatelli, N. S., McIntyre, P. B., Therkildsen, N. O., & Baetscher, D. S. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences* , *117* (51), 32211–32212. doi: 10.1073/pnas.2007421117

Machida, R. J., & Knowlton, N. (2012). PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS ONE* , *7* (9), e46180. doi: 10.1371/journal.pone.0046180

Mariac, C., Vigouroux, Y., Duponchelle, F., Garcia-Davila, C., Nunez, J., Desmarais, E., & Renno, J. F. (2018). Metabarcoding by capture using a single COI probe (MCSP) to identify and quantify fish species in ichthyoplankton swarms. *PLOS ONE* , *13* (9), e0202976. doi: 10.1371/journal.pone.0202976

Marin, A., Serna, J., Robles, C., Ramirez, B., Reyes-Flores, L. E., Zelada-Mazmela, E., ... Alfaro, R. (2018). A glimpse into the genetic diversity of the Peruvian seafood sector: Unveiling species substitution, mislabeling and trade of threatened species. *PLOS ONE* , *13* (11), e0206596. doi: 10.1371/journal.pone.0206596

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* , *17* (1). doi: https://doi.org/10.14806/ej.17.1.200

McInnes, J. C., Alderman, R., Deagle, B. E., Lea, M., Raymond, B., & Jarman, S. N. (2017). Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. *Methods in Ecology and Evolution* , *8* (2), 192–202. doi: 10.1111/2041-210X.12677

McInnes, J. C., Jarman, S. N., Lea, M.-A., Raymond, B., Deagle, B. E., Phillips, R. A., ... Alderman, R. (2017). DNA Metabarcoding as a Marine Conservation and Management Tool: A Circumpolar Examination of Fishery Discards in the Diet of Threatened Albatrosses. *Frontiers in Marine Science* , *4* , 277. doi: 10.3389/fmars.2017.00277

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2019). microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environmental DNA* , *1* (1), 14–25. doi: 10.1002/edn3.11

Meusnier, I., Singer, G. A., Landry, J.-F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* , *9* (1), 214. doi: 10.1186/1471-2164-9-214

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., ... Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science* ,*2* (7), 150088. doi: 10.1098/rsos.150088

Miya, Masaki, Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding: A high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fisheries Science* . doi: 10.1007/s12562-020-01461-x

Mo, W. Y., Man, Y. B., & Wong, M. H. (2018). Use of food waste, fish waste and food processing waste for China's aquaculture industry: Needs and challenge. *Science of The Total Environment* , *613–614* , 635–643. doi: 10.1016/j.scitotenv.2017.08.321

Oksanen, J., F. Guillaume Blanchet, Friendly, M., Roeland Kindt, Pierre Legendre, Dan McGlinn, ... Wagner, H. (2019). Vegan: Community Ecology Package. R package version 2.5-6 (Version 2.5-6).

R Core Team. (2019). *R: A language and environment for statistical computing* . Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ratnasingham, S., & Hebert, P. D. N. (2007). BARCODING: Bold: The Barcode of Life Data System (http://www.barcodinglife.org): BARCODING.*Molecular Ecology Notes* , *7* (3), 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Ripp, F., Krombholz, C., Liu, Y., Weber, M., Schafer, A., Schmidt, B., ... Hankeln, T. (2014). All-Food-Seq (AFS): A quantifiable screen for species in biological samples by deep DNA sequencing. *BMC Genomics* , *15* (1), 639. doi: 10.1186/1471-2164-15-639

Salter, I., Joensen, M., Kristiansen, R., Steingrund, P., & Vestergaard, P. (2019). Environmental DNA concentrations are correlated with regional biomass of Atlantic cod in oceanic waters.*Communications Biology* , *2* (1), 461. doi: 10.1038/s42003-019-0696-8

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated—Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* , *15* (6), 1289–1303. doi: 10.1111/1755-0998.12402

Shaw, J. L. A., Clarke, L. J., Wedderburn, S. D., Barnes, T. C., Weyrich, L. S., & Cooper, A. (2016). Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system.*Biological Conservation* , *197* , 131–138. doi: 10.1016/j.biocon.2016.03.010

Shen, W., & Xiong, J. (2019). TaxonKit: A cross-platform and efficient NCBI taxonomy toolkit. *BioRxiv* . doi: 10.1101/513523

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., & Hajibabaei, M. (2015). A DNA mini-barcoding system for authentication of processed fish products. *Scientific Reports* , *5* (1), 15894. doi: 10.1038/srep15894

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports* , *5* (1), 9687. doi: 10.1038/srep09687

Silva, L. P., Mata, V. A., Lopes, P. B., Pereira, P., Jarman, S. N., Lopes, R. J., & Beja, P. (2019). Advancing the integration of multi-marker metabarcoding data in dietary analysis of trophic generalists. *Molecular Ecology Resources* , *19* (6), 1420–1432. doi: 10.1111/1755-0998.13060

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., & Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: A case study of eDNA metabarcoding seawater. *Scientific Reports* , *9* (1), 5991. doi: 10.1038/s41598-019-42455-9

Smith, D. P., & Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE* , *9* (2), e90234. doi: 10.1371/journal.pone.0090234

Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., ... Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification.*Analytical and Bioanalytical Chemistry* , *408* (17), 4615–4630. doi: 10.1007/s00216-016-9595-8

Stampar, S. N., Broe, M. B., Macrander, J., Reitzel, A. M., Brugler, M. R., & Daly, M. (2019). Linear Mitochondrial Genome in Anthozoa (Cnidaria): A Case Study in Ceriantharia. *Scientific Reports* ,*9* (1), 6094. doi: 10.1038/s41598-019-42621-z

Tacon, A. G. J., & Metian, M. (2008). Global overview on the use of fish meal and fish oil in industrially compounded aquafeeds: Trends and future prospects. *Aquaculture* , *285* (1–4), 146–158. doi: 10.1016/j.aquaculture.2008.08.015

Tange, O. (2011). GNU Parallel: The Command-Line Power Tool. *The USENIX Magazine* , *63* (1), 42–47.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Moller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. *PLoS ONE* ,*7* (8), e41732. doi: 10.1371/journal.pone.0041732

Valdez-Moreno, M., Ivanova, N. V., Elias-Gutierrez, M., Pedersen, S. L., Bessonov, K., & Hebert, P. D. N. (2019). Using eDNA to biomonitor the fish community in a tropical oligotrophic lake. *PLOS ONE* ,*14* (4), e0215505. doi: 10.1371/journal.pone.0215505

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology* , *25* (4), 929–942. doi: 10.1111/mec.13428

Yeo, D., Srivathsan, A., & Meier, R. (2020a). Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification. *Systematic Biology* , syaa014. doi: 10.1093/sysbio/syaa014

Yeo, D., Srivathsan, A., & Meier, R. (2020b). Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification. *Systematic Biology* , *69* (5), 999–1015. doi: 10.1093/sysbio/syaa014

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution* , 2041-210X.13485. doi: 10.1111/2041-210X.13485

Zinger, L., Bonin, A., Alsos, I. G., Balint, M., Bik, H., Boyer, F., ... Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions.*Molecular Ecology* , *28* (8), 1857–1862. doi: 10.1111/mec.15060

Data accessibility

Data and analyses are available at https://github.com/dbaetscher/metabarcoding-primer-portfolio, doi:10.5281/zendo.5706485. Sequencing data files will be uploaded to the NCBI Sequence Read Archive.

Author contributions

D.S.B., N.S.L., P.B.M. and N.O.T. designed the study with input from all authors. D.S.B. and N.S.L. performed laboratory work and analysed data. D.S.B. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Tables

Table 1. Minibarcoding markers included in this study, the expected PCR product size, PCR annealing temperature used for the thermocycling protocol, and reference information that includes the primer sequences and target taxa. Three markers (Teleo, Crust2, and 16Sfish) received too few sequencing reads were excluded from analysis (locus designated with *).

Table 1. Minibarcodes tested for amplifying aquaculture feeds and reference samples.

Locus	Product size	Barcoding gene	Annealing temperature	Reference
AquaF2/C_FishR1	185	COI	51	Valdez-Moreno et al. 2019
AquaF3/C_FishR1	185	COI	50	Valdez-Moreno et al. 2019
FishF2/Shark COI-MINIR	127	COI	52	Fields et al. 2015, Boussarie et al.
nsCOIFo	124	COI	43	Günther et al. 2018
Fish_miniA	129	COI	46	Shokralla et al. 2015
Crust2*	106	COI	55	Fernandes et al. 2017
MiniBar	130	COI	46	Meusnier et al. 2008
Plank_COI	127	COI	52	Berry et al. 2017
Shark474F/FishR1+FishR2	200	COI	52	Cardenosa et al. 2017
16S_Fish*	220	16S	67	McInnes et al. 2017
Crust_16S	170	16S	51	Berry et al. 2017

Ceph_16S	200	16S	55	Berry et al. 2017
L2513/H2714	244	16S	57	Kitano et al. 2007
CEP	206	16S	58	Giusti et al. 2017
16Spc	295	16S	53	Giusti et al. 2017
16Svar	209	16S	58	Giusti et al. 2017
16S-H1	120	16S	61	Horreo et al. 2013
MiFish	185	12S	58	Miya et al. 2015
Teleo*	102	12S	55	Valentini et al 2016
18S_SSU3	170	18S	67	McInnes et al. 2017
18Sn4	180	18S	55	Machida & Knowlton 2012
Short28S	203	28S	59	Vestheim and Jarman 2008

Table 2. Taxa included in full reference (FR) DNA pool. Those taxa also included in the vouchered reference (VR) DNA pool are indicated. Vouchered specimens were obtained from the Cornell University Museum of Vertebrates (CUMV), Kansas University (KU), the Northeastern University Ocean Genome Legacy Center (NOGLC), and the Smithsonian National Museum of Natural History (USNM) collections. Samples obtained as DNA extracts are indicated with an asterix (*). Teleost fishes are ordered alphabetically by family and non-teleosts are listed separately below with their respective group.

Common name	Species	Family	Voucher	Source
Common thresher shark	*Alopias vulpinus*	Alopiidae	yes	CUMV
Sand eel	*Ammodytes dubius*	Ammodytidae	no	
Broad-banded cardinalfish	*Ostorhinchus fasciatus*	Apogonidae	yes	USNM*
Gray triggerfish	*Balistes capriscus*	Balistidae	yes	CUMV
Pacific saury	*Cololabis saira*	Belonidae	no	
Deepbody boarfish	*Antigonia capros*	Caproidae	yes	CUMV
Round scad	*Decapterus punctatus*	Carangidae	yes	CUMV
Shortfin scad	*Decapterus sp.*	Carangidae	no	
Japanese jack mackerel	*Trachurus japonicus*	Carangidae	yes	USNM*
Horse mackerel	*Trachurus lathami*	Carangidae	yes	USNM*
Atlantic horse mackerel	*Trachurus trachurus*	Carangidae	yes	KU
Sandbar shark	*Carcharhinus plumbeus*	Carcharhinidae	yes	CUMV
White sucker	*Catostomus commersonii*	Catostomidae	no	
Small mouth bass	*Micropterus dolomieu*	Centrarchidae	no	
Dorado	*Salminus hilarii*	Characidae	no	
Tilapia	*Oreochromis niloticus*	Cichlidae	no	
Lake Tanganyikan cichlid	*Petrochromis kazumbe*	Cichlidae	no	
Alewife	*Alosa pseudoharengus*	Clupeidae	yes	CUMV
Gulf menhaden	*Brevoortia patronus*	Clupeidae	yes	USNM*
Atlantic Menhaden	*Brevoortia tyrannus*	Clupeidae	yes	CUMV
Lake Tanganyika sardine	*Limnothrissa miodon*	Clupeidae	no	
Atlantic Thread Herring	*Opisthonema oglinum*	Clupeidae	yes	CUMV
Spanish Sardine	*Sardinella aurita*	Clupeidae	yes	CUMV
Goldstripe sardinella	*Sardinella gibbosa*	Clupeidae	yes	USNM
Herring	*Sardinella sp.*	Clupeidae	no	
Sardinops	*Sardinops sagax*	Clupeidae	yes	KU
Conger Eel	*Conger oceanicus*	Congridae	yes	CUMV
Longhorn sculpin	*Myoxocephalus octodecemspinosus*	Cottidae	yes	CUMV
Labeobarbus sp.	*Labeobarbus sp.*	Cyprinidae	no	
Creek chub	*Semotilus atromaculatus*	Cyprinidae	no	

Common name	Species	Family	Voucher	Source
Bluntnose stingray	*Dasyatis say*	Dasyatidae	yes	CUMV
Southern stingray	*Hypanus americanus*	Dasyatidae	yes	CUMV
Round herring	*Etrumeus teres*	Dussumieriidae	yes	USNM*
Striped anchovy	*Anchoa hepsetus*	Engraulidae	yes	CUMV
Bay anchovy	*Anchoa mitchilli*	Engraulidae	yes	CUMV
Silver anchovy	*Engraulis eurystole*	Engraulidae	yes	CUMV
Japanese anchovy	*Engraulis japonicus*	Engraulidae	no	
California anchovy	*Engraulis mordax*	Engraulidae	yes	KU
Atlantic spadefish	*Chaetodipterus faber*	Ephippidae	yes	CUMV
Antarctic krill	*Euphausia superba*	Euphausiidae	yes	NOGLC*
Red cornetfish	*Fistularia petimba*	Fistulariidae	yes	CUMV
Alaska pollock	*Gadus chalcogrammus*	Gadidae	no	
Atlantic cod	*Gadus morhua*	Gadidae	yes	CUMV
Haddock	*Melanogrammus aeglefinus*	Gadidae	yes	CUMV
Bigeye lates	*Lates mariae*	Latidae	no	
Common ponyfish	*Leiognathus equulus*	Leiognathidae	yes	USNM
Monkfish	*Lophius americanus*	Lophiidae	yes	CUMV
Silver Hake	*Merluccius bilinearis*	Merlucciidae	yes	CUMV
Upside-down catfish	*Synodontis irsacae*	Mochokidae	no	
Japanese goatfish	*Upeneus japonicus*	Mullidae	yes	KU
Skinnycheek lanternfish	*Benthosema pterotum*	Myctophidae	yes	KU
Bullnose ray	*Myliobatis freminvillei*	Myliobatidae	yes	CUMV
Capelin	*Mallotus villosus*	Osmeridae	no	
Swai	*Pangasianodon hypophthalmus*	Pangasiidae	no	
Gulf Stream flounder	*Citharichthys arctifrons*	Paralichthyidae	yes	CUMV
Summer flounder	*Paralichthys dentatus*	Paralichthyidae	yes	CUMV
Red hake	*Urophycis chuss*	Phycidae	yes	CUMV
Spotted codling	*Urophycis regia*	Phycidae	yes	CUMV
South American catfish	*Rhamdia quelen*	Pimelodidae	no	
Righteye flounder	*Glyptocephalus cynoglossus*	Pleuronectidae	yes	CUMV
Winter flounder	*Pseudopleuronectes americanus*	Pleuronectidae	yes	CUMV
Eeltail catfish	*Plotosus lineatus*	Plotosidae	yes	USNM*
Atlantic bigeye	*Priacanthus arenatus*	Priacanthidae	yes	CUMV
Coporo	*Prochilodus mariae*	Prochilodontidae	no	
Rosette skate	*Leucoraja garmani*	Rajidae	yes	CUMV
Winter skate	*Leucoraja ocellata*	Rajidae	no	
Atlantic salmon	*Salmo salar*	Salmonidae	no	
Brook trout	*Salvelinus fontinalis*	Salmonidae	no	
American silver perch	*Bairdiella chrysoura*	Sciaenidae	yes	CUMV
Weakfish	*Cynoscion regalis*	Sciaenidae	yes	CUMV
Banded drum	*Larimus fasciatus*	Sciaenidae	yes	CUMV
Spot	*Leiostomus xanthurus*	Sciaenidae	yes	CUMV
Southern kingfish	*Menticirrhus americanus*	Sciaenidae	yes	CUMV
Atlantic croaker	*Micropogonias undulatus*	Sciaenidae	yes	CUMV
Atlantic saury	*Scomberesox saurus*	Scomberesocidae	yes	CUMV
Bullet tuna	*Auxis rochei*	Scombridae	yes	USNM*
Mackerel tuna	*Euthynnus affinis*	Scombridae	yes	USNM*
Indian mackerel	*Rastrelliger kanagurta*	Scombridae	yes	USNM*
Atlantic bonito	*Sarda sarda*	Scombridae	yes	CUMV
Chub mackerel	*Scomber japonicus*	Scombridae	yes	CUMV

Common name	Species	Family	Voucher	Source	
Mackerel	*Scomber scombrus*	Scombridae	no		
King mackerel	*Scomberomorus cavalla*	Scombridae	yes	CUMV	
Atlantic Spanish mackerel	*Scomberomorus maculatus*	Scombridae	yes	CUMV	
Yellowfin tuna	*Thunnus albacares*	Scombridae	no		
Windowpane flounder	*Scophthalmus aquosus*	Scophthalmidae	yes	CUMV	
Blackbelly rosefish	*Helicolenus dactylopterus*	Sebastidae	yes	CUMV	
Acadian redfish	*Sebastes fasciatus*	Sebastidae	yes	CUMV	
Pacific sergestid	*Sergestes similis*	Sergestidae	yes	NOGLC*	
Rock sea bass	*Centropristis philadelphica*	Serranidae	yes	CUMV	
Pinfish	*Lagodon rhomboides*	Sparidae	yes	CUMV	
Scup	*Stenotomus chrysops*	Sparidae	yes	CUMV	
Marbled swamp eel	*Synbranchus marmoratus*	Synbranchidae	no		
Inshore lizardfish	*Synodus foetens*	Synodontidae	yes	CUMV	
Atlantic cutlassfish	*Trichiurus lepturus*	Trichiuridae	yes	CUMV	
Northern sea robin	*Prionotus carolinus*	Triglidae	yes	CUMV	
Blackwing sea robin	*Prionotus rubio*	Triglidae	yes	CUMV	
Mud minnow	*Umbra limi*	Umbridae	no		
Silvery John dory	*Zenopsis conchifer*	Zeidae	yes	CUMV	
John dory	*Zeus faber*	Zeidae	yes	USNM	
Non-teleost taxa					
Common name	Species	Family	Voucher	Source	Group
Common thresher shark	*Alopias vulpinus*	Alopiidae	yes	CUMV	elasmobranch
Chain dogfish	*Scyliorhinus retifer*	Scyliorhinidae	yes	CUMV	elasmobranch
Indian squid	*Loliolus sp.*	Loliginidae	no		cephalopod
Purpleback flying squid	*Sthenoteuthis oualaniensis*	Ommastrephidae	no		cephalopod
Shrimp	*Litopeneaus vannamei*	Penaeidae	no		crustacean

Table 3. Composition of fishmeal, fillers, and binders included in experimental feeds, and the percentages of those constituents added to the feeds. Ingredients for fillers and binders are consistent with products used in aquaculture. Experimental feeds include 0-100% fishmeal. Feeds with 0-25% fishmeal represent the range of fishmeal added to aquaculture feeds. Fishmeal taxa include representatives from families commonly used in feeds (Clupeidae, Engraulidae) and from cultured species (*Salmo salar* , *Oreochromis sp.* ,*Litopeneaus vannamei*).

Hosted file

`image1.emf` available at https://authorea.com/users/449482/articles/547988-optimizing-a-metabarcoding-primer-portfolio-for-species-level-detection-of-taxa-in-complex-mixtures-of-diverse-fishes

Figures

Hosted file

`image2.emf` available at https://authorea.com/users/449482/articles/547988-optimizing-a-metabarcoding-primer-portfolio-for-species-level-detection-of-taxa-in-complex-mixtures-of-diverse-fishes

Figure 1. Accumulation curve for reference species identified by metabarcoding markers. Markers are ordered with the locus that identifies the most reference taxa in the leftmost position followed iteratively by primers that recover the most additional unique taxa. Moving from left to right, each point indicates either the proportion (panel A) or count (panel B) of known species in the full reference DNA pool identified

to the indicated taxonomic rank (color-coded) by the marker indicated and all preceding primers. The proportions in panel A reflect the number of taxa identified out of the total identified by all markers for a particular taxonomic rank. For species assignments, the total number identified was 61, therefore, this number represents 100% of taxa identified to species-level. All 103 taxa are identified to family. The targeted barcoding gene for each marker is noted in parentheses on the x-axis.

Hosted file

`image3.emf` available at https://authorea.com/users/449482/articles/547988-optimizing-a-metabarcoding-primer-portfolio-for-species-level-detection-of-taxa-in-complex-mixtures-of-diverse-fishes

Figure 2. Number of reference species identified by each locus at one- or more-of three taxonomic levels (i.e. all assignments at the species-level are also matches at the genus-level). Species-level taxonomy is only assigned for sequences matching GenBank reference data at >98% identity. Error bars show variation among nine replicates for a given locus. Filled bars show reference taxa identified exclusively (i.e., not by any of the other loci tested) by the locus indicated to the taxonomic rank shown by color.

Hosted file

`image4.emf` available at https://authorea.com/users/449482/articles/547988-optimizing-a-metabarcoding-primer-portfolio-for-species-level-detection-of-taxa-in-complex-mixtures-of-diverse-fishes

Figure 3. Data decontamination steps remove most false positives from the vouchered reference community. Species-level false positives that persist after decontamination are indicated in gray (other), but in some cases, overlap with species that were removed from other primers during decontamination steps (occupancy modeling or dissimilarity filters). Primers are sorted from left to right by number of vouchered reference taxa correctly identified.

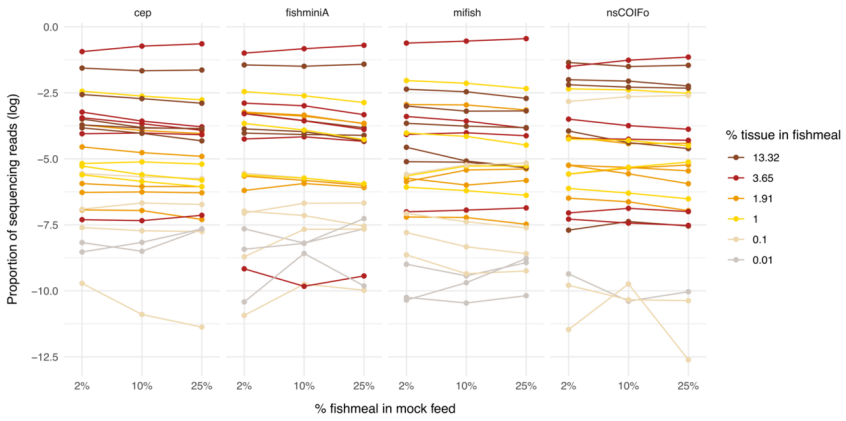

Figure 4. Evidence of amplification bias in quantification of fish taxa in a complex mixture. Proportion of sequencing reads for each reference taxon added to the fish mixture identified to either species or genus. Reads could not be attributed unambiguously for family-level matches. Each point represents a single taxon, which were added to the fishmeal mixture in one of six discrete proportions (% tissue in fishmeal; colors). Lines connect the same taxon across three experimental feeds composed of different proportions fishmeal and filler (10% fishmeal = 90% filler). Taxa that were not detected at all three % fishmeal levels are not displayed (e.g., some taxa at 0.01% tissue).

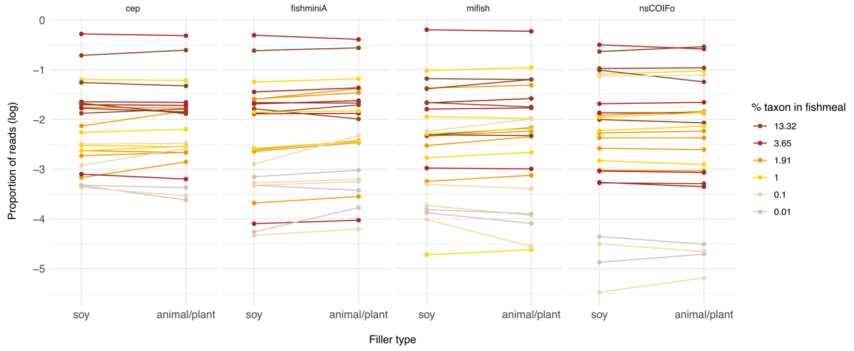

Figure 5. Effect of feed matrix (filler) on proportion of reads recovered for reference taxa. Data are from experimental feed samples with 25% fishmeal and 75% filler. Tissue from known taxa were added to the fishmeal mixture in one of six discrete proportions (% tissue in fishmeal; colors). Soy filler is primarily soy flour with 2% guar gum as a binder. The animal/plant filler includes corn, rice, wheat, and sorghum flour, bloodmeal (from pig, *Sus sp* .), and feathermeal (chicken, *Gallus sp.*) mixed with 2% guar gum for binding (mixture proportions in Table 3). The proportion of sequencing reads (y-axis) is log-scaled.

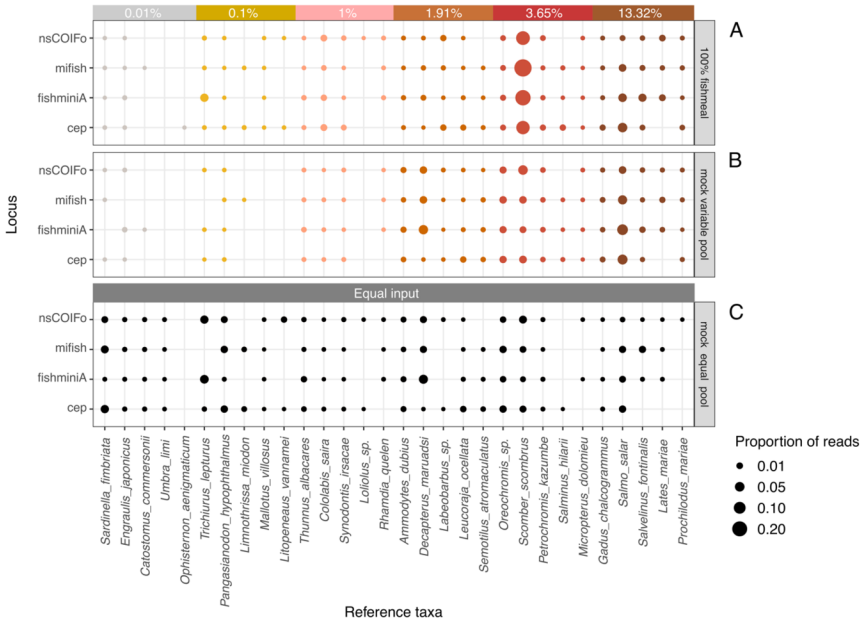

Figure 6. Relationship between tissue added to fishmeal (A) and individual DNA extractions pooled for the same set of reference taxa (B) and between DNA extracts pooled in proportion to the fishmeal mixture (B) or pooled in equal concentration (C). Tissue and DNA concentration % is indicated across the top panel and reference taxa are ordered from least- to most-tissue input. Point size corresponds to the proportion of sequencing reads for a given taxon within each locus (indicated on the y-axis).